

Near-Ideal Behavior of Compressed Sensing Algorithms

Eren Ahsen and M. Vidyasagar *

April 22, 2014

Abstract

In a recent paper [1], see also [2], it is shown that the LASSO algorithm exhibits “near-ideal behavior,” in the following sense: Suppose $y = Az + \eta$ where A satisfies the restricted isometry property (RIP) with a sufficiently small constant, and $\|\eta\|_2 \leq \epsilon$. Then minimizing $\|z\|_1$ subject to $\|y - Az\|_2 \leq \epsilon$ leads to an estimate \hat{x} whose error $\|\hat{x} - x\|_2$ is bounded by a universal constant times the error achieved by an “oracle” that knows the location of the nonzero components of x . In the world of optimization, the LASSO algorithm has been generalized in several directions such as the group LASSO, the sparse group LASSO, either without or with tree-structured overlapping groups, and most recently, the sorted LASSO. In this paper, it is shown that *any algorithm* exhibits near-ideal behavior in the above sense, provided only that (i) the norm used to define the sparsity index is “decomposable,” (ii) the penalty norm that is minimized in an effort to enforce sparsity is “ γ -decomposable,” and (iii) a “compressibility condition” in terms of a group restricted isometry property is satisfied. Specifically, the group LASSO, and the sparse group LASSO (with some permissible overlap in the groups), as well as the sorted ℓ_1 -norm minimization all exhibit near-ideal behavior. Explicit bounds on the residual error are derived that contain previously known results as special cases.

1 Introduction

The field of “compressed sensing” has become very popular in recent years, with an explosion in the number of papers. In the interests of brevity, we

*Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX 75080. Emails: {ahsen,m.vidyasagar}@utdallas.edu. This research was Supported by the National Science Foundation under Awards #1001643 and #1306630, and the Cecil H. & Ida Green Endowment at UT Dallas

refer the reader to two recent papers [2, 3], each of which contains an extensive bibliography. Stated briefly, the core problem in compressed sensing is to approximate a high-dimensional sparse (or nearly sparse) vector x from a small number of linear measurements of x . Though this problem has a very long history (see the discussion in [2] for example), perhaps it is fair to say that much of the recent excitement has arisen from [4], in which it is shown that if x has no more than k nonzero components, then by choosing the matrix A to satisfy a condition known as the restricted isometry property (RIP), it is possible to recover x exactly by minimizing $\|z\|_1$ subject to the constraint that $Az = y = Ax$. In other words, under suitable conditions, among all the preimages of $y = Ax$ under A , the preimage that has minimum ℓ_1 -norm is the sparse signal x itself. The same point is also made in [5]. In case $y = Ax + \eta$ where η is a measurement error and x is either sparse or nearly sparse, one can attempt to recover x by setting

$$\hat{x} := \operatorname{argmin}_{z \in \mathbb{R}^n} \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (1)$$

This algorithm is very closely related to the LASSO algorithm introduced in [6]. Specifically, the only difference between LASSO as in [6] and the problem stated above is that the roles of the objective function and the constraint are reversed. It is shown (see [1, Theorem 1.2] that, under suitable conditions, the residual error $\|\hat{x} - x\|_2$ satisfies an estimate of the form

$$\|\hat{x} - x\|_2 \leq C_0\sigma + C_2\epsilon, \quad (2)$$

where σ is the “sparsity index” of x (defined below), and C_0, C_2 are universal constants that depend only on the matrix A but not x or η . The above bound includes exact signal recovery with noiseless measurements as a special case, and is referred to in [1] as “noisy recovery.”

In a related paper [7], the behavior of the conventional LASSO algorithm is analyzed. Thus one defines

$$\hat{\beta} := \operatorname{argmin}_{z \in \mathbb{R}^n} \|y - Az\|_2^2 + \lambda\|z\|_1,$$

where λ is a Lagrange multiplier.¹ The behavior of the estimate $\hat{\beta}$ is analyzed under the assumption that the vector x is sparse, with the locations of

¹Note that there are some slight departures from the notation used in [7]. Specifically, in [7] there a factor of 1/2 in front of the ℓ_2 -norm, and the Lagrange multiplier λ is weighted by the standard deviation of the measurement noise η . But clearly these are very minor differences.

the nonzero elements chosen in accordance with a very nonrestrictive probabilistic model. It is shown that, under suitable conditions on the matrix A referred to as the “coherence property” and for appropriate choices of the parameter λ , the least squares error of the estimate \hat{x} produced by the standard LASSO algorithm is “nearly ideal” in the following sense: The residual error $\|Ax - A\hat{\beta}\|_2^2$ is bounded by a constant times the error of an “oracle” that knows the locations of the nonzero components of x .

In the world of optimization, the LASSO algorithm has been generalized in several directions, by modifying the ℓ_1 -norm penalty of LASSO to some other norm that induces a prespecified sparsity structure on the solution. One such modification is the Elastic Net (EN) algorithm introduced in [8]; however the penalty in the EN algorithm is not a norm, though it is a convex function. Among the most popular sparsity-inducing penalty norms are the group LASSO [9, 10], referred to hereafter as GL, and the sparse group LASSO [11, 12], referred to hereafter as SGL. Now there are versions of these algorithms that permit the groups to overlap [13, 14]. A recent contribution is to replace the usual ℓ_1 -norm by the “sorted” ℓ_1 -norm [15]. A similar idea is earlier proposed in [16], but in the context of the sorted ℓ_2 -norm.

It is therefore natural to ask whether inequalities analogous to (2) hold when the ℓ_1 -norm in (1) is replaced by other sparsity-inducing norms such as those mentioned in the previous paragraph. More to the point, it would be highly desirable to have a general theory of what properties a norm needs to satisfy, in order that inequalities of the form (2) hold. That is the focus of the present paper. By a slight abuse of terminology, we will refer to the optimization problem formulation in (1) as the “LASSO algorithm,” though as mentioned earlier the roles of the objective function and constraint are reversed in comparison to the conventional LASSO algorithm of [6]. In the same vein, we will refer to the inequality (5) as “near ideal behavior,” and will interpret the results of [1], specifically [1, Theorem 1.2], as saying that the LASSO algorithm exhibits near-ideal behavior. Accordingly, the main objective of this paper is to present a very general result to the effect that *any* compressed sensing algorithm exhibits near-ideal behavior provided it satisfies three conditions:

1. A “compressibility condition” is satisfied, which in the case of LASSO is that the restricted isometry property (RIP) holds with a sufficiently small constant.
2. The approximation norm used to compute the sparsity index of the unknown vector x is “decomposable” as defined subsequently.

3. The penalty norm used to induce the sparsity of the solution, that is, the norm that is minimized, is “ γ -decomposable” as defined subsequently.

It will follow as a consequence of this general result that sorted ℓ_1 -norm minimization, GL, and SGL (without or with tree-structured overlapping groups) all exhibit near-ideal behavior. In addition to the generality of the results established, the method of proof is more direct than that in [1, 2]. Moreover, subject to one additional assumption that always holds in the case of “pure” sparsity (but not in “group” sparsity), the results presented here reduce to those in [1, 2].

In the literature on the behavior of various compressed sensing algorithms, the measurement matrix A is assumed to satisfy a variety of properties, such the restricted isometry property (RIP), the coherence property, etc. Almost all available methods for actually constructing matrices satisfying these properties are *probabilistic*.² In other words, various methods are available which, *with high probability*, result in a matrix A having the desired properties. As a result, the conclusions on near-ideal behavior are also probabilistic. The paper [2], which is in turn an elaboration of [1], is among the few to distinguish clearly between the task of constructing a suitable measurement matrix, and the consequences of doing so. The former relies on probabilistic arguments whereas the latter relies on purely deterministic arguments. Accordingly, in this paper we adopt the same approach as in [1, 2]. This demonstrates the versatility of the approach adopted in [1, 2] to handle a variety of compressed sensing algorithms.

2 Preliminaries

As is by now customary in this area, if $x \in \mathbb{R}^n$, and Λ is a subset of $\mathcal{N} = \{1, \dots, n\}$, the symbol $x_\Lambda \in \mathbb{R}^n$ denotes the vector such that $(x_\Lambda)_i = x_i$ if $i \in \Lambda$, and $(x_\Lambda)_i = 0$ if $i \notin \Lambda$. In other words, x_Λ is obtained from x by replacing x_i by zero whenever $i \notin \Lambda$. Also, as is customary, for a vector $u \in \mathbb{R}^n$, its support set is defined by

$$\text{supp}(u) := \{i : u_i \neq 0\}.$$

Let k be some integer that is fixed throughout the paper. Next we

²Note that it is indeed possible to come up with deterministic algorithms for constructing a measurement matrix with RIP. However, the resulting matrix would be poorly conditioned from a numerical standpoint.

introduce the notion of a group k -sparse set. Some care is required in doing so, as the discussion following the definition shows.

Definition 1. Let $\mathcal{G} = \{G_1, \dots, G_g\}$ be a partition of $\mathcal{N} = \{1, \dots, n\}$, such that $|G_i| \leq k$ for all i . If $S \subseteq \{1, \dots, g\}$, define $G_S := \cup_{i \in S} G_i$. A subset $\Lambda \subseteq \mathcal{N}$ is said to be **S -group k -sparse** for some $S \subseteq \{1, \dots, g\}$ if $\Lambda = G_S$ and $|\Lambda| \leq k$, and **group k -sparse** if it is S -group k -sparse for some set $S \subseteq \{1, \dots, g\}$. A vector $u \in \mathbb{R}^n$ is said to be **group k -sparse** if its support set $\text{supp}(u)$ is contained in a group k -sparse set.

At this point the reader might ask why a set Λ cannot be defined to be group k -sparse if it is a *subset* of some G_S , as opposed to being *exactly equal* to some G_S . The reason is that, if every subset of G_S is also called “group k -sparse,” then in effect all sets of cardinality k or less can be called group k -sparse, thus defeating the very purpose of the definition. To see this, let $\Lambda = \{x_{i_1}, \dots, x_{i_l}\}$, where $l \leq k$, so that $|\Lambda| = l \leq k$. Then, since the sets G_1, \dots, G_g partition the index set \mathcal{N} , for each j there exists a set G_j such that $x_{i_j} \in G_j$. Let $S \subseteq \{1, \dots, g\}$ denote the set consisting of all these indices j . Then $\Lambda \subseteq G_S$. So with this modified definition, there would be no difference between group k -sparsity and conventional sparsity. This is the reason for adopting the above definition. On the other hand, it is easy to see that if $g = n$ and each set G_i consists of exactly one element, then group k -sparsity reduces to conventional k -sparsity. Note also that a vector is defined to be group k -sparse if its support *is contained in*, though not necessarily equal to, a group k -sparse subset of \mathcal{N} .

Suppose $\|\cdot\| : \mathbb{R}^n \rightarrow \mathbb{R}_+$ is some norm. We introduce a couple of notions of decomposability that build upon an earlier definition from [3].

Definition 2. The norm $\|\cdot\|$ is said to be **decomposable** with respect to the partition \mathcal{G} if, whenever $u, v \in \mathbb{R}^n$ are group k -sparse with $\text{supp}(u), \text{supp}(v)$ contained in disjoint group k -sparse sets Λ_u, Λ_v respectively, it is true that

$$\|u + v\| = \|u\| + \|v\|. \quad (3)$$

As pointed out in [3], because $\|\cdot\|$ is a norm, the relationship (3) *always* holds with \leq replacing the equality. Therefore the essence of decomposability is that the bound is tight when the two summands are both group k -sparse vectors with their support sets contained in disjoint group k -sparse subsets of \mathcal{N} . Note that it is not required for (3) to hold for every pair of vectors with disjoint supports, only vectors whose support sets are contained in disjoint group k -sparse subsets of \mathcal{N} . For instance, if Λ is a group k -sparse

set, and u, v have disjoint support sets $\text{supp}(u), \text{supp}(v)$ that are both subsets of Λ , then there is no requirement that (3) hold. We will exploit this flexibility fully in Section 6 where it is shown that the penalty norms in the group LASSO and sparse group LASSO algorithms can be replaced by quite general norms and yet exhibit near-ideal behavior.

As shown in Section 6, the ℓ_1 -norm, the group LASSO and the sparse group LASSO norm are all decomposable. However, the sorted ℓ_1 -norm is not decomposable. To handle that case, we introduce a more general definition.

Definition 3. *The norm $\|\cdot\|$ is γ -decomposable with respect to the partition \mathcal{G} if there exists $\gamma \in (0, 1]$ such that the following is true: whenever $u, v \in \mathbb{R}^n$ are group k -sparse with $\text{supp}(u), \text{supp}(v)$ contained in disjoint group k -sparse sets Λ_u, Λ_v respectively, it is true that*

$$\|u + v\| \geq \|u\| + \gamma\|v\|. \quad (4)$$

Note that if the norm $\|\cdot\|$ is γ -decomposable with $\gamma = 1$, then (4) and the triangle inequality imply that

$$\|u + v\| \geq \|u\| + \|v\| \implies \|u + v\| = \|u\| + \|v\|.$$

Therefore decomposability is the same as γ -decomposability with $\gamma = 1$.

Clearly, if $\|\cdot\|$ is a decomposable norm, then (3) can be applied recursively to show that if $\Lambda_0, \Lambda_1, \dots, \Lambda_s$ are pairwise disjoint group k -sparse sets, and $\text{supp}(u_i) \subseteq \Lambda_i$, then

$$\left\| \sum_{i=0}^s u_i \right\| = \sum_{i=0}^s \|u_i\|, \quad (5)$$

However, such an equality does not hold for γ -decomposable functions unless $\gamma = 1$, which makes the norm decomposable. On the other hand, by repeated application of (4) and noting that $\gamma \leq 1$, we arrive at the following relationship: if $\Lambda_0, \Lambda_1, \dots, \Lambda_s$ are pairwise group k -sparse sets, and $\text{supp}(u_i) \subseteq \Lambda_i$, then

$$\left\| \sum_{i=1}^s u_i \right\| \geq \|u_{\Lambda_1}\| + \gamma \left\| \sum_{i=2}^s u_i \right\|. \quad (6)$$

Equation (5) is somewhat more general than the definition of decomposability given in [3], in that we permit the partitioning of the index set \mathcal{N} into more than two subsets. However, this is a rather minor generalization.³

³There is a little bit of flexibility in [3] in that one can take two orthogonal subspaces that are not exactly orthogonal complements of each other; but we will not belabor this point.

With this preparation we can define the sparsity indices and optimal decompositions. Given an integer k , let GkS denote the collection of all group k -sparse subsets of $\mathcal{N} = \{1, \dots, n\}$, and define

$$\sigma_{k,\mathcal{G}}(x, \|\cdot\|) := \min_{\Lambda \in \text{GkS}} \|x - x_\Lambda\| = \min_{\Lambda \in \text{GkS}} \|x_{\Lambda_0^c}\| \quad (7)$$

to be the **group k -sparsity index** of the vector x with respect to the norm $\|\cdot\|$ and the group structure \mathcal{G} . Since the collection of sets GkS is finite (though it could be huge), we are justified in writing \min instead of \inf . Once we have the definition of the sparsity index, it is natural to define the next notion. Given $x \in \mathbb{R}^n$, and a norm $\|\cdot\|$, we call $\{x_{\Lambda_0}, x_{\Lambda_1}, \dots, x_{\Lambda_s}\}$ an **optimal group k -sparse decomposition** of x if $\Lambda_i \in \text{GkS}$ for $i = 0, \dots, s$, and in addition

$$\|x_{\Lambda_0^c}\| = \min_{\Lambda \in \text{GkS}} \|x - x_\Lambda\|, \|x_{\Lambda_i^c}\| = \min_{\Lambda \in \text{GkS}} \|x - \sum_{j=0}^{i-1} x_{\Lambda_j} - x_\Lambda\|, i = 1, \dots, s.$$

3 Problem Formulation

Throughout we shall make use of three distinct norms:

- $\|\cdot\|_A$, which is a *decomposable norm* that is used to measure the quality of the approximation. Thus, for a vector $x \in \mathbb{R}^n$, the quantity $\sigma_{k,\mathcal{G}}(x, \|\cdot\|_A)$ is the sparsity index used throughout.
- $\|\cdot\|_2$, which is the standard Euclidean or ℓ_2 -norm, and is used to constrain the measurement matrix via the group restricted isometry property (GRIP).
- $\|\cdot\|_P$, which is a γ -*decomposable norm for some $\gamma \in (0, 1]$* , that is minimized to induce a desired sparsity structure on the solution.

The prototypical problem formulation is this: Suppose $x \in \mathbb{R}^n$ is an unknown vector, $A \in \mathbb{R}^{m \times n}$ is a measurement matrix, $y = Ax + \eta$ is a possibly noise-corrupted measurement vector in \mathbb{R}^m , and $\eta \in \mathbb{R}^m$ is the measurement error. It is presumed that $\|\eta\|_2 \leq \epsilon$, where ϵ is a known prior bound. To estimate x from y , we solve the following optimization problem

$$\hat{x} = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \|z\|_P \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (8)$$

Definition 4. The algorithm described in (8) for estimating x is said to be **near ideal** or to show **near-ideal behavior**, if there exist universal constants⁴ D_0 and D_2 that might depend on the matrix A but not on x or η such that⁵

$$\|\hat{x} - x\|_2 \leq D_0 \sigma_{k,\mathcal{G}}(x, \|\cdot\|_A) + D_2 \epsilon. \quad (9)$$

The interpretation of the inequality (9) in this general setting is the same as in [1, 2]. Suppose the vector x is group k -sparse, so that $\sigma_{k,\mathcal{G}}(x, \|\cdot\|_A) = 0$. Then an “oracle” that knows the actual support set of x can approximate x through computing a generalized inverse of the columns of A corresponding to the support of x , and the resulting residual error will be bounded by a multiple of ϵ . Now suppose the algorithm satisfies (9). Then (9) implies that the residual error achieved by the algorithm is bounded by a universal constant times that achieved by an oracle. Proceeding further, (9) also implies that if measurements are noise-free so that $\epsilon = 0$, then the estimate \hat{x} equals x . In other words, the algorithm achieves exact recovery of group k -sparse vectors under noise-free measurements.

The penalty norm $\|\cdot\|_P$ that is minimized in order to determine a sparse approximation to x need not be the same as the approximation norm $\|\cdot\|_A$ used to measure the quality of the resulting approximation. This definition is motivated by the results in [17] in which it is shown that if $\|\cdot\|_P$ is the ℓ_1 -norm but $\|\cdot\|_A$ is the ℓ_2 -norm, then in essence compression is not possible. Of course nothing prevents us from using the same norm for both penalization and approximation. Indeed, except for sorted ℓ_1 -norm minimization, we will take both norms to be the same.

Regarding the use of the ℓ_2 -norm, the theorems proved below are valid for any inner product norm, so long as the partition is orthogonal with respect to the inner product. But the only results available in the literature for constructing measurement matrices with the desired properties are for the ℓ_2 -norm. Hence in the current paper we will use ℓ_2 -norm to constrain the measurement matrix A .

Throughout the paper, we shall be making use of four constants:

$$a := \min_{x \neq 0} \frac{\|x\|_A}{\|x\|_P}, b := \max_{\Lambda \in \text{GkS}} \max_{x_\Lambda \neq 0} \frac{\|x_\Lambda\|_A}{\|x_\Lambda\|_P}, \quad (10)$$

⁴We use the symbols D_0 and D_2 , skipping D_1 , to conform with the notation in [2], where the corresponding constants are denoted by C_0 and C_2 .

⁵The symbol A is unfortunately doing double duty, representing the approximation norm as well as the measurement matrix. After contemplating various options, it was decided to stick to this notation, in the hope that the context would make clear which usage is meant.

$$c := \min_{\Lambda \in \text{GkS}} \min_{x_\Lambda \neq 0} \frac{\|x_\Lambda\|_A}{\|x_\Lambda\|_2}, d := \max_{\Lambda \in \text{GkS}} \max_{x_\Lambda \neq 0} \frac{\|x_\Lambda\|_A}{\|x_\Lambda\|_2}. \quad (11)$$

Note that in the definition of a , the extremum is taken over all nonzero vectors x , whereas in the definitions of b, c, d , the extremum is taken only over all nonzero group k -sparse vectors.

Suppose for instance that $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_1$, which is the approximation as well as penalty norm used in LASSO. Since $|\Lambda| \leq k$ for all $\Lambda \in \text{GkS}$, we have by Schwarz's inequality that

$$\|v\|_1 \leq \sqrt{k} \|v\|_2$$

whenever $\text{supp}(v) \subseteq \Lambda \in \text{GkS}$. In the other direction, we can write

$$v = \sum_{i \in \text{supp}(v)} v_i \mathbf{e}_i,$$

where \mathbf{e}_i is the i -th unit vector. Therefore by the triangle inequality

$$\|v\|_2 \leq \sum_{i \in \text{supp}(v)} \|v_i \mathbf{e}_i\|_2 \leq \sum_{i \in \text{supp}(v)} |v_i| = \|v\|_1.$$

Therefore

$$1 \leq c \leq d \leq \sqrt{k}.$$

Estimates of these constants for other popular norms are given in Section 6.

4 A Known Result

we begin by reprising a known result, as stated in [2], which is a more detailed description of ideas sketched in [1].

Definition 5. Suppose $A \in \mathbb{R}^{m \times n}$. Then we say that A satisfies the **Restricted Isometry Property (RIP)** of order k with constant δ_k if

$$(1 - \delta_k) \|u\|_2^2 \leq \langle u, Au \rangle \leq (1 + \delta_k) \|u\|_2^2, \quad \forall u \in \Sigma_k, \quad (12)$$

where Σ_k denotes the set of all $u \in \mathbb{R}^n$ such that $|\text{supp}(u)| \leq k$.

Theorem 1. ([1, Theorem 1.2], [2, Theorem 1.9]; compare with [7, Theorem 1.4]) Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the RIP of order $2k$ with constant $\delta_{2k} < \sqrt{2} - 1$, and that $y = Ax + \eta$ for some $x \in \mathbb{R}^n$ and $\eta \in \mathbb{R}^m$ with $\|\eta\|_2 \leq \epsilon$. Define

$$\hat{x} = \underset{z \in \mathbb{R}^n}{\text{argmin}} \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (13)$$

Then

$$\|\hat{x} - x\|_2 \leq C_0 \frac{\sigma_k(x, \|\cdot\|_1)}{\sqrt{k}} + C_2 \epsilon, \quad (14)$$

where

$$C_0 = 2 \frac{1 + (\sqrt{2} - 1)\delta_{2k}}{1 - (\sqrt{2} + 1)\delta_{2k}} = 2 \frac{1 + \alpha}{1 - \alpha}, \quad (15)$$

$$C_2 = \frac{4\sqrt{1 + \delta_{2k}}}{1 - (\sqrt{2} + 1)\delta_{2k}} = \frac{4\sqrt{1 + \delta_{2k}}}{(1 - \delta_{2k})(1 - \alpha)}, \quad (16)$$

with

$$\alpha = \frac{\sqrt{2}\delta_{2k}}{1 - \delta_{2k}}. \quad (17)$$

The formulas for C_0, C_2 are written slightly differently from those in [2, Theorem 1.9] but are easily shown to be equivalent to them.

5 Main Results

In this section we present the two main theorems of the paper. The first step is to extend the definition of RIP to group RIP.

Definition 6. A matrix $A \in \mathbb{R}^{m \times n}$ is said to **satisfy the group RIP of order k with constants $\underline{\rho}_k, \bar{\rho}_k$** if

$$0 < \underline{\rho}_k \leq \min_{\Lambda \in \text{GkS}} \min_{\text{supp}(z) \subseteq \Lambda} \frac{\|Az\|_2^2}{\|z\|_2^2}, \quad (18)$$

$$\underline{\rho}_k \geq \max_{\Lambda \in \text{GkS}} \max_{\text{supp}(z) \subseteq \Lambda} \frac{\|Az\|_2^2}{\|z\|_2^2}. \quad (19)$$

We also define

$$\delta_k := \frac{\bar{\rho}_k - \underline{\rho}_k}{2}.$$

Definition 6 shows that the group RIP constants $\underline{\rho}_k$ and $\bar{\rho}_k$ can be a lot closer together than the standard RIP constants in Definition 5, because the various maxima and minima are taken over only group k -sparse sets, and not all subsets of \mathcal{N} of cardinality k . This has implications when probabilistic methods are used to construct the measurement matrix A . We shall return to this topic in Section 7.

Note that in Definition 5, it was assumed that the constants $\underline{\rho}_k$ and $\bar{\rho}_k$ are “symmetric” in that they are of the form $1 - \delta_k, 1 + \delta_k$. We could use a

similar formulation for group RIP as well; but there is no particular reason to do so. The “symmetric” definition is used in Definition 5 to facilitate comparison with earlier results, e.g. those in Theorem 1.

Theorem 2. *Suppose the norm $\|\cdot\|_P$ is γ -decomposable, and that the norm $\|\cdot\|_A$ is decomposable. Define the constants a, b, c, d as in (10) and (11). Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the group RIP property of order $2k$ with constants $(\underline{\rho}_{2k}, \bar{\rho}_{2k})$ respectively, and let $\delta_{2k} = (\bar{\rho}_{2k} - \underline{\rho}_{2k})/2$ as before. Suppose $x \in \mathbb{R}^n$ and that $y = Ax + \eta$ where $\|\eta\|_2 \leq \epsilon$. Define*

$$\hat{x} = \operatorname{argmin}_{z \in \mathbb{R}^n} \|z\|_P \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (20)$$

Suppose the compressibility condition

$$\delta_{2k} < \frac{a\gamma c \underline{\rho}_k}{bd}. \quad (21)$$

is satisfied. Then

$$\|\hat{x} - x\|_A \leq \frac{r_3(r_2 + 1)}{1 - r_1 r_2} \sigma_A + \frac{r_4(r_1 + 1)}{1 - r_1 r_2} \epsilon, \quad (22)$$

$$\|\hat{x} - x\|_2 \leq \frac{r_3(r_2 + 1)}{a(1 - r_1 r_2)} \sigma_A + \frac{r_4(r_1 + 1)}{a(1 - r_1 r_2)} \epsilon, \quad (23)$$

where

$$r_1 = \frac{b}{a\gamma}, r_2 = \frac{\delta_{2k} d}{c \underline{\rho}_k}, r_3 = \frac{b(\gamma + 1)}{a\gamma}, r_4 = \frac{2d\sqrt{\bar{\rho}_k}}{\underline{\rho}_k}, \quad (24)$$

$$\sigma_A := \sigma_{k, \mathcal{G}}(x, \|\cdot\|_A). \quad (25)$$

Before giving the proof of the theorem, we make several observations. Note that, unlike in the earlier result [2, Theorem 1.9], stated here as Theorem 1, the “compressibility condition” (21) involves $\underline{\rho}_k$, and not $\underline{\rho}_{2k}$. Because the set of group k -sparse vectors is a subset of the set of group $2k$ -sparse vectors, it is obvious that $\underline{\rho}_{2k} \leq \underline{\rho}_k \leq \bar{\rho}_k \leq \bar{\rho}_{2k}$. So in this sense (21) is less conservative than any condition that involves replacing $\underline{\rho}_k$ by $\underline{\rho}_{2k}$.

However, in the special case of RIP as opposed to group RIP, a simplification is possible that does not work in more general situations. Specifically, suppose $x \in \mathbb{R}^n$ and that $\{x_{\Lambda_0}, x_{\Lambda_1}, \dots, x_{\Lambda_s}\}$ is an optimal k -sparse (not optimal group k -sparse) decomposition of x with respect to $\|\cdot\|_1$. Then x_{Λ_0} consists of the k largest components of x by magnitude, x_{Λ_1} consists of the next k largest, and so on. One consequence of this is that

$$\min_j |(x_{\Lambda_i})_j| \geq \max_j |(x_{\Lambda_{i+1}})_j|, \quad \forall i,$$

where the minimum on the left is taken over only the nonzero components. Therefore

$$\|x_{\Lambda_{i+1}}\|_2 \leq \sqrt{k} \|x_{\Lambda_{i+1}}\|_\infty \leq \frac{1}{\sqrt{k}} \|x_{\Lambda_i}\|_1. \quad (26)$$

This is the equation just above [1, Equation(10)]. However, when we take optimal *group* k -sparse decompositions, this inequality is no longer valid. For example, suppose $\|\cdot\|_P = \|\cdot\|_1$, let $n = 4, g = 2, k = 2$ and

$$G_1 = \{1, 2\}, G_2 = \{3, 4\}, x = [1 \quad 0.1 \quad 0.6 \quad 0.6]^t.$$

Then it is easy to verify that $s = 2$, and

$$\Lambda_0 = \{3, 4\} = G_2, \Lambda_1 = \{1, 2\} = G_1,$$

$$x_{\Lambda_0} = [0 \quad 0 \quad 0.6 \quad 0.6]^t, x_{\Lambda_1} = [1 \quad 0.1 \quad 0 \quad 0]^t.$$

Here we see that the largest element of x_{Λ_1} is in fact larger than the smallest element of x_{Λ_0} . However, we do not have the freedom to “swap” these elements as they belong to different sets G_i . A more elaborate example is the following: Let $n = 8, g = 4, k = 4$, and

$$x = [0.1 \quad 1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \quad 0.4 \quad 0.7],$$

$$G_1 = \{1\}, G_2 = \{2, 3, 4\}, G_3 = \{5, 6\}, G_4 = \{7, 8\}.$$

Then

$$\Lambda_0 = G_3 \cup G_4, \Lambda_1 = G_1 \cup G_2.$$

Note that x_{G_2} has higher ℓ_1 -norm than any other x_{G_j} . However, since G_2 has cardinality 3, it can only be paired with G_1 , and not with G_3 or G_4 , in order that the cardinality of the union remain less than $k = 4$. And $\|x_{G_1 \cup G_2}\|_1 < \|x_{G_3 \cup G_4}\|_1$.

The proof of Theorem 2 depends on a few preliminary lemmas.

Lemma 1. *Suppose $h \in \mathbb{R}^n$, that $\Lambda_0 \in GkS$ is arbitrary, and let $h_{\Lambda_1}, \dots, h_{\Lambda_s}$ be an optimal group k -sparse decomposition of $h_{\Lambda_0^c}$ with respect to the decomposable approximation norm $\|\cdot\|_A$. Then*

$$\sum_{j=1}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{c} \|h_{\Lambda_0^c}\|_A. \quad (27)$$

Proof: This is a direct consequence of the definition of the constant c and the decomposability of $\|\cdot\|_A$. We reason as follows:

$$\sum_{j=1}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{c} \sum_{j=1}^s \|h_{\Lambda_j}\|_A = \frac{1}{c} \|h_{\Lambda_0^c}\|_A.$$

Lemma 1 is pretty straight-forward, and it is not even necessary for $h_{\Lambda_1}, \dots, h_{\Lambda_s}$ to be an *optimal* group k -sparse decomposition – it can be *any* group k -sparse decomposition, as is easily verified. But it is stated as a separate lemma just to facilitate comparison with [1, Equation (10)], [2, Lemma A.4]. Note that the summation in (27) begins with $j = 1$ and not $j = 2$ as in [1, 2]. If $\|\cdot\|_A$ is the ℓ_1 -norm, then $c = 1$, so the bound in (27) is worse by a factor of \sqrt{k} compared to that in [1, 2]. However, as is evident from the above example, in the case of group k -sparse decompositions, in general it may not be possible to do any better.

Lemma 2 should be compared with [1, Lemma 2.1], [2, Lemma A.3].

Lemma 2. *Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the group RIP of order $2k$ with constants $\underline{\rho}_{2k}$ and $\bar{\rho}_{2k}$, and that u, v are group k -sparse with supports contained in disjoint group k -sparse subsets of \mathcal{N} . Then*

$$|\langle Au, Av \rangle| \leq \delta_{2k} \|u\|_2 \cdot \|v\|_2. \quad (28)$$

Proof: Since we can divide through by $\|u\|_2 \cdot \|v\|_2$, an equivalent statement is the following: If u, v are group k -sparse with supports contained in disjoint group k -sparse subsets of \mathcal{N} , and $\|u\|_2 = \|v\|_2 = 1$, then

$$|\langle Au, Av \rangle| \leq \delta_{2k}.$$

Now the assumptions guarantee that $u \pm v$ are both group $2k$ -sparse. Moreover $u^t v = 0$ since they have disjoint support. Therefore $\|u \pm v\|_2^2 = 2$. So the group RIP implies that

$$2\underline{\rho}_{2k} \leq \|Au \pm Av\|_2^2 \leq 2\bar{\rho}_{2k}.$$

Now the parallelogram identity implies that

$$|\langle Au, Av \rangle| = \left| \frac{\|Au + Av\|_2^2 - \|Au - Av\|_2^2}{4} \right| \leq \frac{2(\bar{\rho}_{2k} - \underline{\rho}_{2k})}{4} = \delta_{2k}.$$

Lemma 3 is the group analog of [2, Lemma 1.3], which is also implicit in [1].

Lemma 3. Suppose $h \in \mathbb{R}^n$, that $\Lambda_0 \in GkS$ is arbitrary, and let $h_{\Lambda_1}, \dots, h_{\Lambda_s}$ be an optimal group k -sparse decomposition of $h_{\Lambda_0^c}$ with respect to the decomposable approximation norm $\|\cdot\|_A$. Then

$$\|h_{\Lambda_0}\|_2 \leq \frac{\delta_{2k}}{c\underline{\rho}_k} \|h_{\Lambda_0^c}\|_A + \frac{\sqrt{\bar{\rho}_k}}{\underline{\rho}_k} \|Ah\|_2. \quad (29)$$

Proof: Note that h_{Λ_0} is group k -sparse. Therefore by the definition of the group RIP property, it follows that

$$\underline{\rho}_k \|h_{\Lambda_0}\|_2^2 \leq \|Ah_{\Lambda_0}\|_2^2 \leq \bar{\rho}_k \|h_{\Lambda_0}\|_2^2.$$

Next, observe that

$$\|Ah_{\Lambda_0}\|_2^2 = \langle Ah_{\Lambda_0}, Ah_{\Lambda_0} \rangle.$$

So we will work on a bound for the right side. We have that

$$\langle Ah_{\Lambda_0}, Ah_{\Lambda_0} \rangle = \langle Ah_{\Lambda_0}, Ah \rangle - \langle Ah_{\Lambda_0}, Ah_{\Lambda_0^c} \rangle.$$

Next by Schwarz's inequality, it follows that

$$\begin{aligned} |\langle Ah_{\Lambda_0}, Ah_{\Lambda_0^c} \rangle| &\leq \left| \sum_{j=1}^s \langle Ah_{\Lambda_0}, Ah_{\Lambda_j} \rangle \right| \\ &\leq \delta_{2k} \|h_{\Lambda_0}\|_2 \sum_{j=1}^s \|h_{\Lambda_j}\|_2 \\ &\leq \frac{\delta_{2k}}{c} \|h_{\Lambda_0}\|_2 \|h_{\Lambda_0^c}\|_A. \end{aligned}$$

Moreover

$$|\langle Ah_{\Lambda_0}, Ah \rangle| \leq \|Ah_{\Lambda_0}\|_2 \cdot \|Ah\|_2 \leq \sqrt{\bar{\rho}_k} \|h_{\Lambda_0}\|_2 \cdot \|Ah\|_2.$$

Combining everything gives

$$\begin{aligned} \underline{\rho}_k \|h_{\Lambda_0}\|_2^2 &\leq \|Ah_{\Lambda_0}\|_2^2 \\ &\leq |\langle Ah_{\Lambda_0}, Ah \rangle| + |\langle Ah_{\Lambda_0}, Ah_{\Lambda_0^c} \rangle| \\ &\leq \frac{\delta_{2k}}{c} \|h_{\Lambda_0}\|_2 \|h_{\Lambda_0^c}\|_A + \sqrt{\bar{\rho}_k} \|h_{\Lambda_0}\|_2 \cdot \|Ah\|_2. \end{aligned}$$

Dividing both sides by $\underline{\rho}_k \|h_{\Lambda_0}\|_2$ leads to (29). \square

Proof of Theorem 2: Define $h := \hat{x} - x$ and note that $x = \hat{x} + h$. Also, since \hat{x} is an optimizer as in (20), it follows that $\|x\|_P \geq \|\hat{x}\|_P = \|x + h\|_P$, which can be written as

$$\|x_{\Lambda_0} + x_{\Lambda_0^c}\|_P \geq \|x_{\Lambda_0} + h_{\Lambda_0} + x_{\Lambda_0^c} + h_{\Lambda_0^c}\|_P.$$

The triangle inequality implies that

$$\|x_{\Lambda_0}\|_P + \|x_{\Lambda_0^c}\|_P \geq \|x_{\Lambda_0} + x_{\Lambda_0^c}\|_P.$$

Combining this with the γ -decomposability of $\|\cdot\|_P$, specifically (6), we can reason as follows:

$$\begin{aligned} \|x_{\Lambda_0}\|_P + \|x_{\Lambda_0^c}\|_P &\geq \|x_{\Lambda_0} + h_{\Lambda_0}\|_P + \gamma\|x_{\Lambda_0^c} + h_{\Lambda_0^c}\|_P \\ &\geq \|x_{\Lambda_0}\|_P - \|h_{\Lambda_0}\|_P - \gamma\|x_{\Lambda_0^c}\|_P + \gamma\|h_{\Lambda_0^c}\|_P. \end{aligned}$$

Cancelling the common term $\|x_{\Lambda_0}\|_P$ leads to

$$\gamma\|h_{\Lambda_0^c}\|_P \leq (\gamma + 1)\|x_{\Lambda_0^c}\|_P + \|h_{\Lambda_0}\|_P. \quad (30)$$

By using the definitions of the constants a, b from (10), the triangle inequality, and the decomposability of the norm $\|\cdot\|_A$, we obtain

$$\begin{aligned} a\gamma\|h_{\Lambda_0^c}\|_A &\leq \gamma\|h_{\Lambda_0^c}\|_P \leq \|h_{\Lambda_0}\|_P + (\gamma + 1)\|x_{\Lambda_0^c}\|_P \\ &\leq \|h_{\Lambda_0}\|_P + (\gamma + 1)\left\|\sum_{j=1}^s x_{\Lambda_j}\right\|_P \\ &\leq b\|h_{\Lambda_0}\|_A + b(\gamma + 1)\sum_{j=1}^s \|x_{\Lambda_j}\|_A \\ &= b\|h_{\Lambda_0}\|_A + b(\gamma + 1)\|x_{\Lambda_0^c}\|_A \\ &= b\|h_{\Lambda_0}\|_A + b(\gamma + 1)\sigma_A, \end{aligned}$$

where σ_A is used as a shorthand for $\sigma_{k,G}(x, \|\cdot\|_A)$. Hence, dividing each side by $a\gamma$ and collecting all terms involving h on the left side, we end up with

$$\|h_{\Lambda_0^c}\|_A - r_1\|h_{\Lambda_0}\|_A \leq r_3\sigma_A, \quad (31)$$

where r_1 and r_3 are given in (24).

We generate another inequality using (29), namely

$$\|h_{\Lambda_0}\|_A \leq d\|h_{\Lambda_0}\|_2 \leq \frac{\delta_{2k}d}{c\rho_k}\|h_{\Lambda_0^c}\|_A + \frac{d\sqrt{\rho_k}}{\rho_k}\|Ah\|_2.$$

Since both x and \hat{x} are feasible for the optimization problem, we have that

$$\|Ah\|_2 = \|A\hat{x} - Ax\|_2 = \|y - Ax - (y - A\hat{x})\|_2 \leq 2\epsilon.$$

Substituting this, and collecting all terms involving h on the left side, leads to

$$-r_2\|h_{\Lambda_0^c}\|_A + \|h_{\Lambda_0}\|_A \leq r_4\epsilon, \quad (32)$$

where r_2 and r_4 are as in (24).

Equations (31) and (32) can be together expressed as a matrix inequality, namely:

$$\begin{bmatrix} 1 & -r_1 \\ -r_2 & 1 \end{bmatrix} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_{\Lambda_0}\|_A \end{bmatrix} \leq \begin{bmatrix} r_3 \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ r_4 \end{bmatrix} \epsilon. \quad (33)$$

Let M denote the 2×2 coefficient matrix on the left side. It has positive diagonal elements and negative off-diagonal elements. So it is easy to see that its inverse will be strictly positive if and only if its determinant is positive. Now the determinant of M is positive if and only if $r_1 r_2 < 1$, which is equivalent to (21). Next, suppose (21) holds. Then we can multiply both sides of (33) by M^{-1} and get

$$\begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_{\Lambda_0}\|_A \end{bmatrix} \leq M^{-1} \left\{ \begin{bmatrix} r_3 \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ r_4 \end{bmatrix} \epsilon \right\}.$$

The triangle inequality now implies that

$$\begin{aligned} \|h\|_A &\leq \|h_{\Lambda_0^c}\|_A + \|h_{\Lambda_0}\|_A \\ &\leq \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_{\Lambda_0}\|_A \end{bmatrix} \\ &\leq \begin{bmatrix} 1 & 1 \end{bmatrix} M^{-1} \left\{ \begin{bmatrix} r_3 \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ r_4 \end{bmatrix} \epsilon \right\}. \end{aligned} \quad (34)$$

Finally, observe that

$$M^{-1} = \frac{1}{1 - r_1 r_2} \begin{bmatrix} 1 & r_1 \\ r_2 & 1 \end{bmatrix}.$$

Substituting into (34) and clearing terms leads to the desired conclusion (22). Equation (23) follows from (22) by using the definition of a . \square

To facilitate comparison with earlier results, we derive an alternate version of Theorem 2, based on an alternate version of Lemma 1. Suppose

$h \in \mathbb{R}^n$, that $\Lambda_0 \in GkS$ is arbitrary, and let $h_{\Lambda_1}, \dots, h_{\Lambda_s}$ be an optimal group k -sparse decomposition of $h_{\Lambda_0^c}$ with respect to the approximation norm $\|\cdot\|_A$. Suppose there exists a constant f such that

$$\sum_{j=2}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{f} \|h_{\Lambda_0^c}\|_A. \quad (35)$$

Note that the summation on the left side of (35) begins with $j = 2$ and not $j = 1$ as in (27). For instance, if we are studying conventional sparsity, wherein $g = n$ and each group consists of a singleton, it is known that f can be taken as \sqrt{k} . See [1, Equation(11)], [2, Lemma A.4]. With this assumption we can state another bound.

Theorem 3. *Suppose $A \in \mathbb{R}^{m \times n}$ satisfies the group RIP property of order $2k$ with constants $(\underline{\rho}_{2k}, \bar{\rho}_{2k})$ respectively, and let $\delta_{2k} = (\bar{\rho}_{2k} - \underline{\rho}_{2k})/2$ as before. Suppose $x \in \mathbb{R}^n$ and that $y = Ax + \eta$ where $\|\eta\|_2 \leq \epsilon$. Define*

$$\hat{x} = \underset{z \in \mathbb{R}^n}{\operatorname{argmin}} \|z\|_P \text{ s.t. } \|y - Az\|_2 \leq \epsilon. \quad (36)$$

Suppose that

$$\delta_{2k} < \frac{f \underline{\rho}_{2k}}{\sqrt{2} r_1 d}. \quad (37)$$

Then

$$\|\hat{x} - x\|_2 \leq \frac{r_3(g + 1/f)}{1 - w} \sigma_A + \frac{r_5(1 + (r_1 d)/f)}{1 - w} \epsilon, \quad (38)$$

where

$$g := \frac{\sqrt{2} \delta_{2k}}{f \underline{\rho}_{2k}}, w = r_1 d g, r_5 := \frac{2 \sqrt{\bar{\rho}_k}}{\underline{\rho}_k}, \quad (39)$$

σ_A are defined in (25), c is defined in (11).

If we compare the compressibility conditions (37) and (21), we see that $\underline{\rho}_k$ is replaced by $\underline{\rho}_{2k}$ which is in general smaller, and there is an extra factor of $\sqrt{2}$ in the denominator. Both of these are in favor of the bound in (21). However, the constant c is replaced by the constant f . In the case of “pure” sparsity, $c = 1$ whereas $f = \sqrt{k}$, which is a substantial advantage in favor of (37). In the case of group sparsity, it is unclear whether there is any advantage to (37). As we shall see below, the main advantage of (37) and of Theorem 3 is that it reduces to Theorem 1 in the case of pure sparsity.

To prove the theorem, we need an alternate version of Lemma 3, which is entirely analogous to [2, Lemma 1.3].

Lemma 4. Suppose $h \in \mathbb{R}^n$, that $\Lambda_0 \in GkS$ is arbitrary, and let $h_{\Lambda_1}, \dots, h_{\Lambda_s}$ be an optimal group k -sparse decomposition of $h_{\Lambda_0^c}$ with respect to the approximation norm $\|\cdot\|_A$. Define $\Lambda = \Lambda_0 \cup \Lambda_1$. Then

$$\|h_\Lambda\|_2 \leq \frac{\sqrt{2}\delta_{2k}}{f\rho_{2k}}\|h_{\Lambda_0^c}\|_A + \frac{\sqrt{\rho_{2k}}}{\rho_{2k}}\|Ah\|_2. \quad (40)$$

The proof is a fairly straight-forward modification of that of Lemma 3 and step by step the same as that of [2, Lemma 1.3]. But it is presented in detail, in the interest of completeness.

Proof of Lemma 4: Note that h_Λ is group $2k$ -sparse. Therefore by the definition of the group RIP property, it follows that

$$\rho_{2k}\|h_\Lambda\|_2^2 \leq \|Ah_\Lambda\|_2^2 \leq \bar{\rho}_{2k}\|h_\Lambda\|_2^2.$$

Next, observe that

$$\|Ah_\Lambda\|_2^2 = \langle Ah_\Lambda, Ah_\Lambda \rangle.$$

So we will work on a bound for the right side. Note that

$$\langle Ah_\Lambda, Ah_\Lambda \rangle = \langle Ah_\Lambda, Ah \rangle - \langle Ah_\Lambda, Ah_{\Lambda^c} \rangle.$$

Next by (35) and Schwarz's inequality, it follows that

$$\begin{aligned} |\langle Ah_\Lambda, Ah_{\Lambda^c} \rangle| &\leq \left| \sum_{i=0}^1 \sum_{j=2}^s \langle Ah_{\Lambda_i}, Ah_{\Lambda_j} \rangle \right| \\ &\leq \delta_{2k}[\|h_{\Lambda_0}\|_2 + \|h_{\Lambda_1}\|_2] \sum_{j=2}^s \|h_{\Lambda_j}\|_2 \\ &\leq \frac{\sqrt{2}\delta_{2k}}{f}\|h_\Lambda\|_2\|h_{\Lambda^c}\|_A. \end{aligned}$$

In the above, we use the known inequality

$$\|h_{\Lambda_0}\|_2 + \|h_{\Lambda_1}\|_2 \leq \sqrt{2}\|h_{\Lambda_0} + h_{\Lambda_1}\|_2 = \sqrt{2}\|h_\Lambda\|_2,$$

because h_{Λ_0} and h_{Λ_1} are orthogonal. Next

$$|\langle Ah_\Lambda, Ah \rangle| \leq \|Ah_\Lambda\|_2 \cdot \|Ah\|_2 \leq \sqrt{\rho_{2k}}\|h_{\Lambda_0}\|_2 \cdot \|Ah\|_2.$$

Combining everything gives

$$\begin{aligned} \rho_{2k}\|h_\Lambda\|_2^2 &\leq \|Ah_\Lambda\|_2^2 \\ &\leq |\langle Ah_\Lambda, Ah \rangle| + |\langle Ah_\Lambda, Ah_{\Lambda^c} \rangle| \\ &\leq \frac{\sqrt{2}\delta_{2k}}{f}\|h_\Lambda\|_2\|h_{\Lambda^c}\|_A + \sqrt{\rho_{2k}}\|h_{\Lambda_0}\|_2 \cdot \|Ah\|_2. \end{aligned}$$

Dividing both sides by $\underline{\rho}_{2k}\|h_\Lambda\|_2$ leads to (40). \square

Proof of Theorem 3: As before, the optimality of \hat{x} and the γ -decomposability of the penalty norm $\|\cdot\|_P$ imply that

$$\|h_{\Lambda_0^c}\|_A \leq r_3\|x_{\Lambda_0^c}\|_A + r_1\|h_{\Lambda_0}\|_A,$$

where r_1 and r_3 are given in (24); this is the same equation as (31). Now rewrite this as

$$\|h_{\Lambda_0^c}\|_A \leq r_3\sigma_A + r_1\|h_{\Lambda_0}\|_A \leq r_3\sigma_A + r_1d\|h_{\Lambda_0}\|_2 \leq r_3\sigma_A + r_1d\|h_\Lambda\|_2, \quad (41)$$

because Λ_0 is a subset of Λ . Next, the feasibility of both x and \hat{x} implies that $\|Ah\|_2 \leq 2\epsilon$. Therefore (40) now becomes

$$\|h_\Lambda\|_2 \leq \frac{\sqrt{2}\delta_{2k}}{f\underline{\rho}_{2k}}\|h_{\Lambda_0^c}\|_A + \frac{2\sqrt{\underline{\rho}_{2k}}\epsilon}{\underline{\rho}_{2k}}. \quad (42)$$

The inequalities (41) and (42) can be written as a matrix inequality, namely

$$\begin{bmatrix} 1 & -r_1d \\ -g & 1 \end{bmatrix} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_\Lambda\|_2 \end{bmatrix} \leq \begin{bmatrix} r_3 \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ 1 \end{bmatrix} r_5\epsilon,$$

where g and r_5 are defined in (39). The coefficient matrix on the left side has a strictly positive inverse if its determinant is positive. Moreover, the determinant is just $1 - r_1dg = 1 - w$, where w is also defined in (39). So the compressibility condition is $w < 1$, which is clearly equivalent to (37). Moreover, if $w < 1$, then one can infer from the above matrix inequality that

$$\begin{aligned} \begin{bmatrix} \|h_{\Lambda_0^c}\|_A \\ \|h_\Lambda\|_2 \end{bmatrix} &\leq \frac{1}{1-w} \begin{bmatrix} 1 & r_1d \\ g & 1 \end{bmatrix} \left\{ \begin{bmatrix} r_3 \\ 0 \end{bmatrix} \sigma_A + \begin{bmatrix} 0 \\ r_5 \end{bmatrix} \epsilon \right\} \\ &= \frac{1}{1-w} \left\{ \begin{bmatrix} 1 \\ g \end{bmatrix} r_3\sigma_A + \begin{bmatrix} r_1d \\ 1 \end{bmatrix} r_5\epsilon \right\}. \end{aligned}$$

Now by (35),

$$\|h_{\Lambda^c}\|_2 \leq \sum_{j=2}^s \|h_{\Lambda_j}\|_2 \leq \frac{1}{f} \|h_{\Lambda_0^c}\|_A.$$

Therefore, since $h = h_\Lambda + h_{\Lambda^c}$, the triangle inequality implies that

$$\begin{aligned}
\|h\|_2 &\leq \|h_{\Lambda^c}\|_2 + \|h_\Lambda\|_2 \\
&\leq \frac{1}{f}\|h_{\Lambda_0^c}\|_A + \|h_\Lambda\|_2 \\
&\leq \frac{1}{1-w} \begin{bmatrix} 1/f & 1 \end{bmatrix} \left\{ \begin{bmatrix} 1 \\ g \end{bmatrix} r_3 \sigma_A + \begin{bmatrix} r_1 d \\ 1 \end{bmatrix} r_5 \epsilon \right\} \\
&= \frac{1}{1-w} [r_3(g + 1/f)\sigma_A + (1 + (r_1 d)/f)r_5\epsilon].
\end{aligned}$$

6 Special Cases

In this section it is shown that Theorem 2 is general enough to encompass several algorithms such as group LASSO, sparse group LASSO, either without overlapping groups or with groups that overlap but have a tree structure, and sorted ℓ_1 -norm minimization. Estimates for the constants a , b , c and d in (10), (11) that appear in this theorem are derived for each of these algorithms. Then Theorem 3 is applied to pure sparsity, and Theorem 1 is derived as a special case. Note that our use of terminology such as LASSO, group LASSO is consistent with that in [1, 2], but inconsistent with that in, for instance, [7], in the sense that the roles of the objective function and constraint are interchanged.

First we begin with various versions of the group LASSO algorithm. As before, let $\mathcal{G} = \{G_1, \dots, G_g\}$ be a partition of the index set $\mathcal{N} = \{1, \dots, n\}$. Let $\|\cdot\|_i : \mathbb{R}^{|G_i|} \rightarrow \mathbb{R}_+$ be *any* norm, and define the corresponding norm on \mathbb{R}^n by

$$\|x\|_A = \sum_{i=1}^g \|x_{G_i}\|_i.$$

Then it is easy to see that the above norm is decomposable. The exact nature of the individual norms $\|\cdot\|_i$ is entirely irrelevant. So if we define $\|\cdot\|_A$ as above and $\|\cdot\|_P = \|\cdot\|_A$, then it follows from Theorem 2 that the corresponding algorithm exhibits near-ideal behavior, provided an appropriate compressibility condition holds.

By defining the individual norms $\|\cdot\|_i$ appropriately, it is possible to recover the group LASSO [9, 10], the sparse group LASSO [11, 12], and the overlapping sparse group LASSO with tree-structured norms [13, 14]. The group LASSO norm is defined by

$$\|z\|_{\text{GL}} := \sum_{i=1}^g \|z_{G_i}\|_2. \quad (43)$$

This corresponds to the choice $\|\cdot\|_i = \|\cdot\|_2$. Note that some authors use $\|z_{G_i}\|_2/\sqrt{|G_i|}$ instead of just $\|z_{G_i}\|_2$. The adjustments needed to handle to this variation are elementary and are left to the reader. The sparse group LASSO norm is defined by

$$\|z\|_{\text{SGL},\mu} := \sum_{i=1}^g [(1-\mu)\|z_{G_i}\|_1 + \mu\|z_{G_i}\|_2]. \quad (44)$$

This corresponds to the choice

$$\|z_{G_i}\|_i = (1-\mu)\|z_{G_i}\|_1 + \mu\|z_{G_i}\|_2.$$

Next let us turn our attention to the case of “overlapping” groups with tree structure, as defined in [13, 14]. In this case there are sets $\mathcal{N}_1, \dots, \mathcal{N}_l$, each of which is a subset of \mathcal{N} , that satisfy the condition

$$\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset \implies (\mathcal{N}_i \subseteq \mathcal{N}_j \text{ or } \mathcal{N}_j \subseteq \mathcal{N}_i). \quad (45)$$

Though it is possible for some of these sets to overlap, the condition above implies that the collection of sets $\mathcal{N}_1, \dots, \mathcal{N}_l$ can be renumbered with double indices as \mathcal{S}_{ij} , and arranged in chains of the form

$$\mathcal{S}_{11} \subseteq \dots \subseteq \mathcal{S}_{1n_1}, \dots, \mathcal{S}_{s1} \subseteq \dots \subseteq \mathcal{S}_{sn_s},$$

where the “maximal” sets \mathcal{S}_{in_i} must also satisfy (46). Therefore, given two maximal sets $\mathcal{S}_{in_i}, \mathcal{S}_{jn_j}$, either they must be the same or they must be disjoint, because it is not possible for one of them to be a subset of the other. This shows that the maximal sets \mathcal{S}_{in_i} are pairwise disjoint once the duplicates are removed, and together span the total feature set $\mathcal{N} = \{1, \dots, n\}$. Thus, in a collection of tree-structured sets, the highest level sets do not overlap! Let g denote the number of distinct maximal sets, and define $G_i = \mathcal{S}_{in_i}$ for $i = 1, \dots, g$. Then $\{G_1, \dots, G_g\}$ is a partition of \mathcal{N} , and each \mathcal{N}_j is a subset of some G_i . Now the authors of [13, 14] define a combination of ℓ_1 - and ℓ_2 -norms on the various subsets $\mathcal{N}_1, \dots, \mathcal{N}_l$ and add them up. However, since each set \mathcal{N}_j is a subset of some G_i , a linear combination of norms on all sets \mathcal{N}_j is just a sum of norms on the various G_i and is therefore decomposable. Indeed there is no reason to restrict oneself to convex combinations of ℓ_1 - and ℓ_2 -norms on the subsets \mathcal{N}_j . No matter what norms are defined on these subsets, the resulting summation is still some linear combination of norms on the sets G_i , and as a result the overall norm on \mathbb{R}^n is decomposable.

Thus to summarize, the group LASSO norm, the sparse group LASSO norm, and the penalty norms defined in [13, 14] are all decomposable. Therefore the minimization of any of these norms leads to near-ideal behavior under appropriate compressibility conditions.

Next we turn to the sorted ℓ_1 -norm introduced in [15], which is defined as follows: Suppose $\lambda_1 \geq \dots \geq \lambda_n > 0$, and let $\boldsymbol{\lambda} \in \mathbb{R}_+^n$ denote the vector $[\lambda_i, i = 1, \dots, n]$. Given a vector $x \in \mathbb{R}^n$, let (x) denote its “sorted” version; in other words, (x) is a permutation of x , with the property that $|(x)_i| \geq |(x)_j|$ if $i > j$. Then the **sorted ℓ_1 -norm** of x is defined as

$$\|x\|_{\text{SL1}} := \sum_{i=1}^n \lambda_i |(x)_i|. \quad (46)$$

It is not difficult to verify that the sorted ℓ_1 -norm is *not* decomposable unless all λ_i are equal, in which case it is just (a multiple of) the ℓ_1 -norm. However, it is γ -decomposable with $\gamma = \lambda_n/\lambda_1$. To see this, suppose $u, v \in \mathbb{R}^n$ with disjoint supports, and let $s_1 = |\text{supp}(u)|$, $s_2 = |\text{supp}(v)|$. Since we have defined γ as λ_n/λ_1 , and the λ_i are nonincreasing, it is easy to see that $\lambda_j/\lambda_i \geq \gamma$ whenever $1 \leq i < j \leq n$. In particular we have that $\lambda_{s_1+i} \geq \gamma \lambda_i$ for all i between 1 and $n - s_1$. Also, since the vector $(u + v)$ consists of the elements of $u + v$ arranged in decreasing order of magnitude, and u, v have disjoint support sets, it is evident that

$$|(u + v)_i| \geq |(u + v)_j| \text{ if } i < j,$$

$$|(u + v)_i| \geq \max\{|(u)_i|, |(v)_i|\}, i = 1, \dots, \min\{s_1, s_2\}.$$

Therefore it follows from the definition of the sorted ℓ_1 -norm that

$$\begin{aligned} \|u + v\|_{\text{SL1}} &= \sum_{i=1}^{s_1+s_2} \lambda_i |(u + v)_i| \\ &= \sum_{i=1}^{s_1} \lambda_i |(u + v)_i| + \sum_{i=1}^{s_2} \lambda_{s_1+i} |(u + v)_{s_1+i}| \\ &\geq \sum_{i=1}^{s_1} \lambda_i |(u + v)_i| + \gamma \sum_{i=1}^{s_2} \lambda_i |(u + v)_i| \\ &\geq \sum_{i=1}^{s_1} \lambda_i |(u)_i| + \gamma \sum_{i=1}^{s_2} \lambda_i |(v)_i| \\ &= \|u\|_{\text{SL1}} + \gamma \|v\|_{\text{SL1}}. \end{aligned}$$

In the above line of reasoning, we have majorized λ_{s_1+i} by $\gamma\lambda_i$ and $|(u+v)_{s_1+i}|$ by $|(u+v)_i|$. The last step establishes that the sorted ℓ_1 -norm is γ -decomposable with $\gamma = \lambda_n/\lambda_1$. Therefore it follows from Theorem 2 that sorted ℓ_1 -norm minimization leads to near-ideal behavior under appropriate compressibility conditions.

In the remainder of the section, we determine the constants a , b , c and d in (10), (11) that appear in Theorem 2 for various penalty norms. We consider two situations: In the first case, $\|\cdot\|_A$ is the group LASSO or sparse group LASSO norm. Since these norms are decomposable, we can as well take $\|\cdot\|_P = \|\cdot\|_A$, which makes $a = b = 1$, and we need only to determine the constants c and d . In the second case, $\|\cdot\|_A$ is the ℓ_1 -norm and $\|\cdot\|_P$ is the sorted ℓ_1 -norm.

Consider now the group LASSO and the sparse group LASSO algorithms. First, let $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_{\text{GL}}$. Then since $\|\cdot\|_P = \|\cdot\|_A$, we have $a = b = 1$. To calculate c and d , define l_{\min} to be the smallest cardinality of any G_i , and define $s_{\max} := \lfloor k/l_{\min} \rfloor$. Now suppose that $\Lambda \in \text{GkS}$. Specifically, suppose $\Lambda = G_{i_1} \cup \dots \cup G_{i_s}$. Then clearly

$$\|z_\Lambda\|_{\text{GL}} = \sum_{j=1}^s \|z_{G_{i_j}}\|_2,$$

while

$$\|z_\Lambda\|_2 = \left(\sum_{j=1}^s \|z_{G_{i_j}}\|_2^2 \right)^{1/2}.$$

Thus, if we define the s -dimensional vector $v \in \mathbb{R}_+^s$ by

$$v = [\|z_{G_{i_j}}\|_2, j = 1, \dots, s],$$

then

$$\|z_\Lambda\|_{\text{GL}} = \|v\|_1, \|z_\Lambda\|_2 = \|v\|_2.$$

Now, by earlier discussion, we know that

$$\|v\|_2 \leq \|v\|_1 \leq \sqrt{s}\|v\|_2.$$

Moreover, it is clear that the integer s , denoting the number of distinct sets that make up Λ , cannot exceed s_{\max} . This shows that

$$1 \leq c_{\text{GL}} \leq d_{\text{GL}} \leq \sqrt{s_{\max}}. \quad (47)$$

Next we analyze the sparse group LASSO algorithm, wherein $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_{\text{SGL},\mu}$. Then again we have that $a = b = 1$. To calculate c and

d , suppose $\Lambda = G_{i_1} \cup \dots \cup G_{i_s}$. Let l_{\max} denote the largest cardinality of any G_i . Then

$$\|z_{G_{i_j}}\|_2 \leq \|z_{G_{i_j}}\|_1 \leq \sqrt{l_{\max}} \|z_{G_{i_j}}\|_2,$$

whence

$$\sum_{j=1}^s \|z_{G_{i_j}}\|_2 \leq \sum_{j=1}^s \|z_{G_{i_j}}\|_1 \leq \sqrt{l_{\max}} \sum_{j=1}^s \|z_{G_{i_j}}\|_2. \quad (48)$$

Combining (47) and (48) leads to

$$\|z_\Lambda\|_2 \leq \|z_\Lambda\|_{\text{SGL},\mu} \leq [(1-\mu)\sqrt{l_{\max}} + \mu\sqrt{s_{\max}}] \|z_\Lambda\|_2.$$

Therefore

$$1 \leq c_{\text{SGL},\mu} \leq d_{\text{SGL},\mu} \leq (1-\mu)\sqrt{l_{\max}} + \mu\sqrt{s_{\max}}. \quad (49)$$

Finally consider the case where $\|\cdot\|_A$ is the ℓ_1 -norm and $\|\cdot\|_P$ equals the sorted ℓ_1 -norm. Then it is easy to see that

$$\lambda_n \|x\|_1 \leq \|x\|_P \leq \lambda_1 \|x\|_1.$$

Therefore we can take

$$\frac{1}{\lambda_1} = a_{\text{SL}1} \leq b_{\text{SL}1} = \frac{1}{\lambda_n}.$$

It is already known that $1 \leq c \leq d \leq \sqrt{k}$.

We conclude this section by studying conventional (not group) k -sparsity, and showing that the bounds in Theorem 1 can be obtained from Theorem 3. First of all we have $\|\cdot\|_A = \|\cdot\|_P = \|\cdot\|_1$. Since $\|\cdot\|_P$ is decomposable, we have $\gamma = 1$ and $\|\cdot\|_A = \|\cdot\|_P$ implies that $a = b = 1$. Therefore, $r_1 = 1$ and $r_3 = 2$, where r_1 and r_3 are given in (24). It is known from [1, Equation (10)], [2, Lemma A.4] that (35) holds with $f = \sqrt{k}$. since we are now dealing with conventional ℓ_1 - and ℓ_2 -norms, we can take $c = 1$ and $d = \sqrt{k}$. Moreover, the formulation of the RIP in (12) means that

$$\underline{\rho}_{2k} = 1 - \delta_{2k}, \bar{\rho}_{2k} = 1 + \delta_{2k}.$$

With these values, the constants g and w in (39) become

$$g = \frac{\sqrt{k}\delta_{2k}}{\sqrt{k}(1-\delta_{2k})} = \frac{\alpha}{\sqrt{k}}, w = r_1 dg = \frac{\sqrt{k}\delta_{2k}}{(1-\delta_{2k})} = \alpha,$$

where the constant α is defined in (17). Therefore the two constants appearing in (38) become

$$\frac{r_3(g + 1/f)}{1 - w} = \frac{2}{1 - \alpha} \frac{1}{\sqrt{k}} (\alpha + 1) = \frac{2}{\sqrt{k}} \frac{1 + \alpha}{1 - \alpha} = \frac{C_0}{\sqrt{k}},$$

and

$$\frac{r_5(1 + (r_1 d)/f)}{1 - w} = \frac{2r_5}{1 - \alpha} = \frac{4\sqrt{1 + \delta_{2k}}}{(1 - \alpha)(1 - \delta_{2k})} = C_2,$$

where C_0, C_2 are defined in (15) and (16) respectively. Therefore we have rederived (14) using the bounds in Theorem 3, specifically (38).

7 Sample Sizes

In this section we apply the method of [18] to derive the size of the measurement matrix in order for it to satisfy the group RIP with a specified level of confidence. Specifically, we consider the following problem: Suppose the matrix A is constructed by drawing mn i.i.d. samples of a random variable \mathcal{X} , so that $a_{ij} = \phi_{ij}(\omega)$, where ω denotes the element of the probability space that generates the samples, and i, j denote the row and column indices as usual. Suppose that integers n (the size of the vector x) and k (the sparsity of x though in unknown locations) are specified, along with numbers δ denoting the group RIP constant and ζ denoting a confidence parameter. The objective is to derive lower bounds on the integer m such that, with probability no smaller than $1 - \zeta$, the $m \times n$ matrix A satisfies the group RIP condition of order k with constant no larger than δ .

Specifically, suppose that \mathcal{X} is a zero-mean random variable with variance $1/m$, and let $\phi_{ij}(\omega), 1 \leq i \leq m, 1 \leq j \leq n$ be i.i.d. copies of \mathcal{X} , where ω denotes the element of the sample space. Let $\Phi(\omega) \in \mathbb{R}^{m \times n}$ denote the resulting random $m \times n$ matrix. Then it is easy to verify that

$$E(\|\Phi(\omega)v\|_2^2) = \|v\|_2^2.$$

Now suppose there exists a constant $c_0(\epsilon)$ depending only on ϵ such that

$$\Pr\{|\|\Phi(\omega)v\|_2^2 - \|v\|_2^2| > \epsilon\} \leq 2 \exp[-mc_0(\epsilon)]. \quad (50)$$

Such bounds are established with $c_0(\epsilon) = \epsilon^2/4 - \epsilon^3/6$ if \mathcal{X} is Gaussian with zero mean and variance $1/m$ in [18], and for Bernoulli random variables in [19].

Now [18, Lemma 6.1] states that if $T \subseteq \mathcal{N}$ and $|T| \leq k$, then

$$(1 - \delta)\|x_T\|_2^2 \leq \|\Phi(\omega)x_T\|_2^2 \leq (1 + \delta)\|x_T\|_2^2, \quad \forall x \in \mathbb{R}^n \quad (51)$$

with probability $1 - \alpha'$, where

$$\alpha' \leq 2 \left(\frac{12}{\delta} \right)^k e^{-mc_0(\delta/2)}.$$

Now observe that, for Gaussian or Bernoulli variables,

$$c_0(\epsilon) = \frac{\epsilon^2}{4} - \frac{\epsilon^3}{6} \geq \frac{\epsilon^2}{8}, \quad \forall \epsilon \leq 0.75. \quad (52)$$

Therefore for Gaussian or Bernoulli variables, it follows that

$$\alpha' \leq 2 \left(\frac{12}{\delta} \right)^k e^{-m\delta^2/32}, \quad \forall \delta \leq 0.75. \quad (53)$$

Equations (51) and (53) hold for *any one choice* of the subset $T \subseteq \mathcal{N}$, no matter what it might be. Therefore, if $\mathcal{J} \subseteq 2^{\mathcal{N}}$ is *any arbitrary collection* of subsets of \mathcal{N} where each set $T \in \mathcal{J}$ has cardinality $\leq k$, then by the union bounds formula it follows that

$$(1 - \delta)\|x_T\|_2^2 \leq \|\Phi(\omega)x_T\|_2^2 \leq (1 + \delta)\|x_T\|_2^2, \quad \forall T \in \mathcal{J}, \quad \forall x \in \mathbb{R}^n \quad (54)$$

with probability $1 - \alpha$, where $\alpha = |\mathcal{J}|\alpha'$, that is

$$\alpha \leq 2|\mathcal{J}| \left(\frac{12}{\delta} \right)^k e^{-m\delta^2/32} \quad \forall \delta \leq 0.75. \quad (55)$$

Therefore, given any collection of subsets $\mathcal{J} \subseteq 2^{\mathcal{N}}$ such that each set $T \in \mathcal{J}$ has cardinality $\leq k$, all we have to do is to ensure that the following inequality is satisfied:

$$2|\mathcal{J}| \left(\frac{12}{\delta} \right)^k e^{-m\delta^2/32} \leq \zeta. \quad (56)$$

If every quantity is fixed except m , we can rewrite the above inequality in terms of m , as follows:

$$m \geq \frac{32}{\delta^2} \left[k \log \frac{12}{\delta} + \log |\mathcal{J}| + \log \frac{2}{\zeta} \right]. \quad (57)$$

To apply this approach to constructing matrices that have the group RIP, the only thing needed is to find an upper bound on the size of the collection GkS, the collection of all group k -sparse sets. This is easily done using Sauer’s lemma,⁶ which is well-known in statistical learning theory, and states that for integers $1 \leq d \leq n$, it is true that

$$\sum_{i=0}^d \binom{n}{i} \leq \left(\frac{en}{d}\right)^d, \quad (58)$$

where e denotes the base of the natural logarithm. Sauer’s lemma can be found in many places, out of which [20, Theorem 4.1] is just one reference. To apply Sauer’s lemma to the problem of group sparsity, as before let $\mathcal{G} = \{G_1, \dots, G_s\}$ denote a partition of \mathcal{N} . Let l_{\min} denote $\min_i |G_i|$, and let s_{\max} denote the largest integer s such that there exists a $\Lambda \in \text{GkS}$ of the form $\Lambda = G_{i_1} \cup \dots \cup G_{i_s}$. Clearly $s_{\max} \leq k/l_{\min}$. Therefore

$$|\text{GkS}| \leq \sum_{i=0}^{s_{\max}} \binom{g}{i} \leq \left(\frac{eg}{s_{\max}}\right)^{s_{\max}}.$$

Therefore, given integers n, m, k and numbers $\delta \in (0, 0.75], \zeta \in (0, 1)$, the measurement matrix $A = \Phi(\omega)$ satisfies the group RIP of order k with coefficients $\underline{\rho}_k \geq 1 - \delta$ and $\bar{\rho}_k \leq 1 + \delta$, with probability at least $1 - \zeta$, provided the following inequality holds, which is obtained from (57) by replacing $|\mathcal{J}|$ by the bound from Sauer’s lemma:

$$m_{\text{GS}} \geq \frac{32}{\delta^2} \left[k \log \frac{12}{\delta} + s_{\max}(1 + \log g - \log s_{\max}) + \log \frac{2}{\zeta} \right]. \quad (59)$$

Here the subscript “GS” denotes “group sparsity.” On the other hand, for “pure” sparsity, GkS becomes the set of all k -sparse subsets of \mathcal{N} , whose cardinality is bounded (again by Sauer’s lemma) by $(en/k)^k$. Therefore

$$m_{\text{S}} \geq \frac{32}{\delta^2} \left[k \log \frac{12}{\delta} + k(1 + \log n - \log k) + \log \frac{2}{\zeta} \right]. \quad (60)$$

Here the subscript “S” denotes “sparsity.”

These formulas contain some good news and some bad news. The first piece of good news is that the quantity ζ enters through the logarithm, so

⁶ Actually Sauer’s lemma is more general than this, and bounds the number of subsets of a set of cardinality n that can be generated by intersecting with a collection of VC-dimension d .

that m increases very slowly as we decrease ζ . This is consistent with the well-known maxim in statistical learning theory that “confidence is cheaper than accuracy.” Another piece of good news is that asymptotically we have

$$m_S \sim \frac{k \log n}{\delta^2}.$$

However, the *really bad news* is that in order for the logarithmic effect to start showing up, n has to be really huge.

To illustrate this last comment, let us apply the bounds in (59) and (60) to typical numbers from microarray experiments in cancer biology. Accordingly, we take $n = 20,000$, which is roughly equal to the number of genes in the human body and the number of measured quantities in a typical experiment, and we take $k = 20$, which is a typical number of key biomarkers that we hope will explain most of the observations. Since $\delta \leq \sqrt{2} - 1$ is the compressibility condition for pure sparsity, we take $\delta = 1/4 = 0.25$. We partition the set of 20,000 measurements into $g = 6,000$ sets representing the number of pathways that we wish to study, and we take $l_{\min} = 4$, meaning that the shortest pathway of interest has four genes. Therefore we can take $s_{\max} = \lfloor k/l_{\min} \rfloor = 5$. Finally, let us take $\zeta = 10^{-6}$. With these numbers, it is readily verified that

$$m_S = 71,286, m_{GS} = 47,960.$$

In other words, both values of m are *larger than n* ! Therefore one can only conclude that the known bounds for m are too coarse to be of practical use at least in computational biology, though perhaps they might be of use in other applications where n is a few orders of magnitude larger. For example, with $n = 10^6$, $k = 50$, $g = 3 \times 10^5$ and $s = 5$, the choices $\delta = 0.25$ and $\zeta = 10^{-6}$ lead to

$$m_S = 385,660, m_{GS} = 137,260.$$

Thus, though this time both m_S and m_{GS} are smaller than n , the difference is not substantial.

As to whether group sparsity offers any advantage over pure sparsity, the evidence is not conclusive. A comparison of (59) and (60) shows that m_S grows as $(k \log n)/\delta^2$, whereas m_{GS} grows as $(k + s_{\max} \log g)/\delta^2$. This observation is also made in [10]. Again, “asymptotically” the term involving s_{\max} will dominate k as n becomes very large. However, since s_{\max} is multiplied by $\log g$, g would have to be truly enormous in order for group sparsity to lead to substantially smaller values for m than with pure sparsity.

The contents of this section can be summarized very simply. While very general and universal bounds are available for RIP that can be easily adapted to group RIP, these bounds are basically worthless at the scale of numbers that appear in computational biology.

8 Conclusions

In this paper we have presented a very general theory that allows us to conclude that a compressed sensing algorithm shows near-ideal behavior under extremely general conditions. The results presented here show that not just LASSO, but also group LASSO, sparse group LASSO, and their variants with tree-structured overlapping groups, and sorted ℓ_1 -norm minimization, all show near-ideal behavior. Along the way, we have introduced a notion of γ -decomposability, which generalizes the notion of decomposability. The next logical step would be to remove the restriction of tree structure on overlapping groups.

References

- [1] E. Candès, The restricted isometry property and its implications for compressed sensing, *Comptes rendus de l'Académie des Sciences, Série I* 346 (2008) 589–592.
- [2] M. A. Davenport, M. F. Duarte, Y. C. Eldar, G. Kutyniok, Introduction to compressed sensing, in: Y. C. Eldar, G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012, pp. 1–68.
- [3] S. Negabhan, P. Ravikumar, M. J. Wainwright, B. Yu, A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers, *Statistical Science* 27(4) (December 2012) 538–557.
- [4] E. J. Candès, T. Tao, Decoding by linear programming, *IEEE Transactions on Information Theory* 51 (2005) 4203–4215.
- [5] D. Donoho, For most large underdetermined systems of linear equations, the minimal ℓ_1 -norm solution is also the sparsest solution, *Communications in Pure and Applied Mathematics* 59(6) (2006) 797–829.
- [6] R. Tibshirani, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society* 58(1).

- [7] E. J. Candès, Y. Plan, Near ideal model selection by ℓ_1 minimization, *The Annals of Statistics* 37(5A) (2009) 2145–2177.
- [8] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. Royal Stat. Soc. B* 67 (2005) 301–320.
- [9] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B* 68 (2006) 49–67.
- [10] J. Huang, T. Zhang, The benefit of group sparsity, *The Annals of Statistics* 38(4) (2010) 1978–2004.
- [11] J. Friedman, T. Hastie, R. Tibshirani, A note on the group lasso and sparse group lasso, <http://www-stat.stanford.edu/tibs/ftp/sparse-grlasso.pdf> (2010).
- [12] N. Simon, J. Friedman, T. Hastie, R. Tibshirani, A sparse group lasso, <http://www-stat.stanford.edu/nsimon/SGLpaper.pdf> (2012).
- [13] R. Jenetton, J. Mairal, G. Obozinski, F. Bach, Proximal methods for hierarchical sparse coding, *Journal of Machine Learning Research* 12 (2011) 2297–2334.
- [14] G. Obozinski, L. Jacob, J.-P. Vert, Group lasso with overlaps: The latest group lasso approach, *arxiv* (2011) 1110.0413.
- [15] M. Bogdan, E. van den Berg, W. Su, E. J. Candès, Statistical estimation and testing via the sorted ℓ_1 -norm, <http://statweb.stanford.edu/candes/papers/SortedL1.pdf> (2013).
- [16] I. Daubechies, R. Devore, M. Fornasier, C. S. Güntürk, Iteratively reweighted least squares minimization for sparse recovery, *Communications on Pure and Applied Mathematics* 63(1) (2010) 1–38.
- [17] A. Cohen, Wolfgang, Dahmen, R. Devore, Compressed sensing and best k -term approximation, *Journal of the American Mathematical Society* 22(1) (January 2009) 211–231.
- [18] R. Baraniuk, M. Davenport, R. Devore, M. Wakin, A simple proof of the restricted isometry property for random matrices, *Constructive Approximation* 28 (2008) 253–263.

- [19] D. Achlioptas, Database-friendly random projections: Johnson-Lindenstraus with binary coins, *Journal of Computer and System Sciences* 66 (2003) 671–687.
- [20] M. Vidyasagar, *Learning and Generalization: With Applications to Neural Networks and Control Systems*, Springer-Verlag, London, 2003.